Assessing Probability of Ancestry Using Simple Sequence Repeat Profiles: Applications to Maize Hybrids and Inbreds

Donald A. Berry,[*,1] Jon D. Seltzer,[†] Chongqing Xie,[‡] Deanne L. Wright[‡] and J. Stephen C. Smith[‡]

[*]The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, [†]Third Wave Technologies, Inc., Madison, Wisconsin 53719 and [‡]Pioneer Hi-Bred International, Inc., Johnston, Iowa 50131

Manuscript received July 24, 2001
Accepted for publication March 11, 2002

ABSTRACT

Determination of parentage is fundamental to the study of biology and to applications such as the identification of pedigrees. Limitations to studies of parentage have stemmed from the use of an insufficient number of hypervariable loci and mismatches of alleles that can be caused by mutation or by laboratory error and that can generate false exclusions. Furthermore, most studies of parentage have been limited to comparisons of small numbers of specific parent-progeny triplets thereby precluding large-scale surveys of candidates where there may be no prior knowledge of parentage. We present an algorithm that can determine probability of parentage in circumstances where there is no prior knowledge of pedigree and that is robust in the face of missing data or mistyped data. We present data from 54 maize hybrids and 586 maize inbreds that were profiled using 195 SSR loci including simulations of additional levels of missing and mistyped data to demonstrate the utility and flexibility of this algorithm.

DETERMINATION of parentage is fundamental to the study of reproductive and behavioral biology. The increasing availability of highly discriminant genetic markers for many diverse species provides the potential to uniquely characterize individuals at numerous loci and to unambiguously resolve parentage where genealogical relationships are unknown, in error, or in dispute.

Identification of parent-progeny relationships in wild populations of animals and plants provides insights into the success of various reproductive strategies (Ellstrand 1984; Smouse and Meagher 1994; Alderson *et al.* 1999) and has allowed for the implementation of management programs to conserve genetic diversity (Miller 1975; Rannala and Mountain 1997). The association of pedigree with physical appearance or performance in domesticated animals and plants allows parents that have contributed favorable alleles for desirable traits through selective breeding programs to be identified (Bowers and Meredith 1997; Sefc *et al.* 1998; Vankan and Faddy 1999). These applications of associative genetics facilitate further progress in genetic improvement through breeding. Establishment of parentage is also useful to secure legal rights of guardianship in humans, to help protect intellectual property in plant varieties, to validate breed pedigrees of domesticated animals, to protect stocks of fish, and to identify (Gotz and Thaller 1998; Primmer *et al.* 2000; White *et al.* 2000).

Most studies of pedigree have utilized exclusion analysis where the molecular marker genotypes of either one or a restricted number of potential triplets of offspring and putative parents are compared. Often the identity of the mother is not in question; the maternal profile is subtracted from that of the offspring and the deduced paternal profile is then compared with candidate father genotypes (Ellstrand 1984; Hamrick and Schnabel 1985). Individuals who could not have contributed the paternal genotype are excluded; the remainder are possible parents. Nonpaternity in humans is generally declared only on the basis of exclusions exhibited by at least two unlinked and independent loci. This criterion of exclusion reduces the likelihood of a false declaration of nonpaternity on the basis of marker results that are actually due to mutation within the phylogeny. Bein *et al.* (1998) show that evidence of nonpaternity should require exclusions at loci on different chromosomes to avoid erroneous conclusions that would be made due to nondisjunction at meiosis leading to uniparental inheritance. A requirement for at least three independent exclusions to declare nonpaternity in humans has also been instituted (Gunn *et al.* 1997). In studies of natural populations of animals or plants where numerous parent-progeny triplets are examined it is usual to accept

[illegible] circumstances where the putative father is deceased but his parents are still alive ([illegible] *et al.*

---

[1] [illegible] Texas M. D. Anderson Cancer [illegible] Houston, TX [illegible]

1991; Bockel *et al.* 1992). Chakraborty *et al.* (1994) demonstrate that paternity can be determined in cases where the mother is unavailable for testing. Lang *et al.* (1993) partially reconstructed the DNA profile of a missing crocodile parent using profiles of the mother and progeny.

Chakraborty *et al.* (1988) and Smouse and Meagher (1994) report that reliance upon exclusion alone has usually failed to unambiguously resolve paternity. Limitations have stemmed from the use of an insufficient number of independent hypervariable loci. Other statistical methods are therefore required to calculate the likelihood of paternity for each nonexcluded male (Berry and Geisser 1986; Meagher 1986; Meagher and Thompson 1986; Thompson and Meagher 1987; Devlin *et al.* 1988; Berry 1991). Marshall *et al.* (1998) draw attention to the quality of data that is encountered practically in genotypic surveys. Maternal genetic data may or may not be available, data may be absent for some candidate males, data may be missing for some loci in some individuals, null alleles exist, and typing errors occur. Reconstructing or validating the pedigrees of varieties of cultivated plants often provides additional challenges because their phylogenies can reveal apparent exclusions that masquerade as non-Mendelian inheritance. For example, apparent exclusions can occur in circumstances where an individual is used as a parent prior to completion of the inbreeding process. The development of parent and progeny then continue on parallel but separate tracks thereby allowing the possibility that alleles that are subsequently lost through inbreeding in the parent can still become fixed in the progeny. It is also possible to create many offspring from a single mating and to use the same parent repeatedly in "backcrossing." Therefore, many individual inbred lines, varieties, or hybrids can be highly related. In consequence, there are numerous (and often very similar) pedigrees. The effective number of marker loci that can discriminate between alternate pedigrees is proportionally reduced as parents are increasingly related. Consequently, inbred lines can be more similar to one or more sister or other inbreds than those inbreds are to one or both of their parents.

It has not been usual to search among hundreds of individuals to identify the most probable maternal and paternal candidates for a specific progeny. Most studies of parentage are in circumstances where there is *a priori* information for at least one of the parents (usually the maternal parent). Limited availability of marker loci and the lack of very high-throughput genotyping systems offering inexpensive datapoint costs may have focused

reproductive behavior renders identification of the paternal parent difficult or impossible. Examples include those undertaken on birds that practice brood parasitism (Alderson *et al.* 1999) or extra-pair copulation (Wetton *et al.* 1992) or on species such as the wombat that are difficult to observe in the wild (Taylor *et al.* 1997).

Two circumstances favor a revised approach to the statistical analysis of pedigree. First, molecular marker technologies are rapidly developing and will allow numerous loci to be typed for thousands of individuals rapidly and inexpensively. A greater number and diversity of larger-scale studies of pedigree can be expected within the plant and animal kingdoms including individuals in which there is no prior knowledge of pedigree. A larger number of markers mean a greater chance for errors. Therefore, the second circumstance follows: Procedures that are efficient and robust in the face of apparent exclusions, missing data, and laboratory error are required.

The purpose of this article is to describe and evaluate a methodology that can be used to quantify the probability of parentage of hybrid genotypes. We focus on parentage because it is the primary focus of published literature and it is the easiest level of ancestry to understand. The method is robust in the face of mutation, pseudo-non-Mendelian inheritance (apparent exclusions) due to residual heterozygosity in parental seed sources, missing data, and laboratory error. The methodology has a number of advantages: (i) It can accommodate large datasets of possible ancestors (hundreds of inbreds or hybrids each profiled by >100 marker loci), (ii) it does not require prior knowledge about either parent of the hybrid of interest, (iii) it does not require independence of the markers, and (iv) it can successfully discriminate between many highly related and genetically similar genotypes. We demonstrate the effectiveness of this approach to identify inbred parents of maize (*Zea mays* L.) hybrids using simple sequence repeat (SSR) marker profiles for 54 maize hybrids together with their parental and grandparental genotypes included among a total of 586 inbred lines. The methodology is applicable to the investigation of parentage for all progeny developed from parental mating without subsequent generations of inbreeding.

MATERIALS AND METHODS

Algorithm: Consider an index hybrid whose parentage is unknown or in dispute. Inbreds in an available database are possible ancestors of the hybrid. The objective is to find the probabilities of closest ancestry for each inbred on the basis of information from SSRs from the index hybrid and the others. The [illegible]

[illegible] that maximizes [illegible] repeating this for all pairs of inbreds in the database.

The basis of the algorithm is Bayes' rule (*e.g.*, BERRY 1991, 1996). Let $P(i, j|\text{SSRs})$ stand for the (posterior) probability that *i* and *j* are ancestors of the index hybrid given the information from the various SSRs. Let $P(i, j)$ stand for the unconditional (or prior) probability of the same event. Finally $P(\text{SSRs}|i, j)$ is the probability of observing the various SSR results if in fact *i* and *j* are ancestors. Bayes' rule says

$$P(i, j|\text{SSRs}) = P(\text{SSRs}|i, j) \times P(i, j)/\sum[P(\text{SSRs}|u, v) \times P(u, v)],$$

where the sum in the denominator is over all pairs of inbreds, indexed by *u* and *v*. $P(\text{SSRs}|i, j) \times P(i, j)$ is one of the terms in the denominator. (To compute the denominator in the above expression, fix a particular order to the inbreds in the database and take $u < v$ in expressions involving the pair (u, v). If there are 586 inbreds, for example, then the number of pairs and the number of terms in the denominator is $586(587)/2 = 171{,}991$.) Inbreds *i* and *j* may be parents or grandparents or other types of relations or bear no relationship at all to the hybrid. If there are more than two ancestors in the database, such as both parents and all four grandparents, then the possible pairs involving these ancestors will generally have the highest posterior probabilities. If the hybrid's true parents are in the database, then as a pair they will typically have the highest overall posterior probability. If both *i* and *j* happen to be related to one particular parent of the hybrid, then as a pair their posterior probability will be low because they will not usually account for many of the alleles that are contributed by the other parent of the hybrid.

We will make the "no-prior-information" assumption that $P(u, v)$ is the same for all pairs (u, v). This implies that this factor is cancelled from both numerator and denominator in the above expression, giving:

$$P(i, j|\text{SSRs}) = P(\text{SSRs}|i, j)/\sum P(\text{SSRs}|u, v).$$

The problem is then to calculate a typical $P(\text{SSRs}|i, j)$. Assume inbreds *i* and *j* are both ancestors. We calculate the probability of observing the resulting hybrid under this assumption. We make no assumptions about relationships among the various inbreds. Other possible ancestors will be considered implicitly in the calculation by allowing their alleles to be introduced through breedings with *i* and *j*. However, the nature of such breedings is not specified. Suppose inbred *i*'s alleles are (a, b). Each descendant of inbred *i* receives one of these two alleles or not. An immediate descendant receives one with probability 1 (barring mutations). A second generation descendant receives one of them with probability 0.5. And so on. Since degree of ancestry (if any) is unknown, we label the actual probability of passing on one of these alleles to be *P*. Similarly, an allele from inbred *j* has been passed down to the hybrid or not, and the probability of the former is *P*. In the following, *P* will be taken to equal 0.50, although we will also consider $P = 0.99$ in some of the calculations.

Assuming $P = 0.50$ is consistent with the closest ancestors in the database being grandparents. However, we are not interested in grandparents *per se*. If the closest ancestors in the database were parents, then as indicated above *P* should equal 1 (ignoring mutations and laboratory errors). Our primary concern is when the parents are not in the database. In this case *P* is no greater than 0.50. Assuming $P = 0.50$ is robust over the middle range of possible values of *P*. One way in which it is robust is if there may be mutations and laboratory

[illegible]
tors in the database are more remote than grandparents, then

are likely to be identified because they will usually have the fewest mismatches of the lines considered.

When *i* and *j* are ancestors there are four possibilities: (1) The alleles of both inbreds *i* and *j* were passed to the hybrid (2) inbred *i* came through but not inbred *j*, (3) inbred *j* came through but not inbred *i*, and (4) neither inbred came through. Assuming independence, these have respective probabilities P^2, $P(1 - P)$, $P(1 - P)$, $(1 - P)^2$. In the case $P = 0.50$, all of these probabilities equal 0.25.

An instance of the law of total probability (Sec. 5.3, BERRY 1996) is that the probability of observing a hybrid's alleles is the average of the conditional probability of this event given the above four cases. The simplest of the four cases is the first possibility: Assuming the hybrid's alleles are passed down directly from both inbreds, the probability of observing the hybrid's genotype is either 1 or 0 depending on whether the hybrid shares both inbreds' alleles. (It is especially easy when both inbreds are homozygous.) The other three cases require an assumption regarding the possibility that an inbred's allele is not passed to the hybrid but is interrupted by a mutation, a laboratory error, or intervening breeding. We regard such an allele as being selected from all known alleles with probability 1/(number of alleles), where the number of alleles is the total number of alleles known to exist at the locus in question. An alternative approach would be to use the allelic proportions that are present in the database (or in another database). However, the lines in the database may not be randomly selected from any population. For example, a line that has been highly used in breeding would have many derivative lines in the database, in which case the frequencies of its alleles will be artificially inflated. Assuming equal probabilities for the various alleles at a given locus is robust in the sense that it is not affected by adding and dropping lines from the database.

There are many cases to consider when computing the probability of observing a hybrid's alleles, depending on the zygosity of the hybrid and the inbreds, and allowing for the possibility of missing alleles or "extra alleles" in the assessment of the hybrid and inbred genotypes. These possibilities are too numerous to list. Instead we give three simple examples. All the examples have homozygous inbreds, the most common case. And each of the three hybrids has two alleles, again the most common case. We suppose that the measured alleles for three SSRs and a particular trio of hybrid and ancestor inbreds are as we have indicated in Table 1.

For SSR 1 there are three known alleles, one in addition to alleles *a* and *b* that are listed for the three lines (hybrid, inbred *i*, and inbred *j*) in Table 1. For SSR 2 and SSR 3 there are two known alleles in addition to those listed. The calculations in the right half of Table 1 will now be explained. Implicit in calculating $P(\text{SSR}|i, j)$ is the assumption—required in both the numerator and denominator of Bayes' rule—that inbreds *i* and *j* are ancestors of the hybrid. Consider SSR 1. In case 1 above, both ancestors' alleles (as measured by the laboratory process) are assumed to pass to the index hybrid, and so in this case the hybrid is necessarily *ab*. The probability of observing the actual hybrid's genotype is 1 for case 1, as shown in Table 1. In case 2, we assume that inbred *i*'s allele passes to the hybrid but inbred *j*'s does not. Indeed, the hybrid has an *a* allele. The probability of observing a *b* as the other allele is 1/(number of alleles) = 1/3, as shown in Table 1. Case 3 is similar. In case 4, neither ancestor allele is passed

[illegible]
2 there is some evidence against pair [illegible] being ancestors

TABLE 1

Probability of observing a hybrid's alleles using three sample SSRs and four possible combinations (cases) of alleles passed, assuming that inbreds *i* and *j* are ancestors of the hybrid

					Probability of observing the hybrid's genotype				
SSR	No. of alleles	Hybrid	Inbred *i*	Inbred *j*	Case 1 *i, j*	Case 2 *i, not j*	Case 3 *not i, j*	Case 4 *not i, not j*	Overall probability $P(SSR\|i, j)$
1	3	*ab*	*aa*	*Bb*	1	1/3	1/3	2/9	17/36
2	5	*bd*	*bb*	*Cc*	0	1/5	0	2/25	7/100
3	6	*ab*	*cc*	*Dd*	0	0	0	2/36	2/144

SSR, simple sequence repeat marker profile.

but it is not conclusive. For SSR 3 there is even less evidence favoring pair (i, j). It would not take many SSRs with evidence similar to that for SSR 3 to essentially rule out this pair—provided that other pairs are not similarly inconsistent.

To find the overall $P(SSRs|i, j)$, multiply the individual $P(SSR|i, j)$ over the various SSRs. There are purely computational issues to address. Each $P(SSR|i, j)$ is a number between 0 and 1. When there are a great many SSRs, the product of these numbers will be vanishingly small. To lessen problems with computational underflow, for each SSR we multiply $P(SSR|u, v)$ by the same constant for each pair (u, v): the inverse of the largest possible such probability. For example, since 17/36 is the largest probability for a heterozygous hybrid at an SSR having three alleles (as is the case for SSR 1 in Table 1), we multiply all factors $P(SSR1|u, v)$ by 36/17. To eliminate remaining problems with underflow, we do calculations using logarithms (adding instead of multiplying) and take antilogs at the end.

The probability $P(SSR|u, v)$ is calculated for all (u, v) pairs and summed over all possible pairings in the database, including that for the inbred pair under consideration: (i, j). This gives the denominator in the expression for $P(i, j|SSRs)$.

To determine the probability that any particular inbred, say inbred i, is the closest ancestor of the index hybrid, sum $P(SSR|i, v)$ over all inbreds v with $v \neq i$. Call this $P(i|SSRs)$. The maximum of $P(i|SSRs)$ for any inbred i is 1. But since there is one closest ancestor on each side of the family, the sum of $P(i|SSRs)$ over all inbreds i is 2. If there is a particular pair (i, j) for which $P(i, j|SSRs)$ is close to 1 then both $P(i|SSRs)$ and $P(j|SSRs)$ separately will be close to 1.

SSR data: DNA was extracted from 54 maize hybrids and from 586 maize inbreds. All of the hybrids and most inbreds are proprietary products of Pioneer Hi-Bred International; some important publicly bred inbred lines were also included. The inbred parents and grandparents of each hybrid were included within the set of inbreds. Other inbreds that were genotyped include many that are highly related by pedigree to parents and grandparents of the hybrids. The hybrids were chosen because each has a pedigree that is known to us and collectively they represent a broad array of diversity of maize germplasm that is currently grown in the United States ranging from early to late maturity.

A total of 195 SSR loci were used in this study following

[illegible]

SSR [illegible] the following [illegible] parentheses [illegible] on individual maize chromosomes as follows: 1 (35), 2 (2[illegible]), 3 (22), 4 (20), 5 [illegible], 6 (9), 7 (6), 8 (18), 9 (12), and 10 (14); 17 SSR loci have not yet been mapped. The correlations among the loci are unknown and are irrelevant for our methodology.

Sequence data for primers that allow many of these (and other) SSR loci to be assayed are available at website http://www.agron.missouri.edu. All primers were designed to anneal and amplify under a single set of conditions for PCR in 10-μl reactions. Genomic DNA (10 ng) was amplified in 1.5 mm $MgCl_2$, 50 mm KCl, 10 mm Tris-Cl (pH 8.3) using 0.3 units AmpliTaq Gold DNA polymerase (PE Corporation) oligonucleotide primer pairs (one primer of each pair was fluorescently labeled) at 0.17 μm and 0.2 mm dNTPs. This mixture was incubated at 95° for 10 min (hot start); amplified using 45 cycles of denaturation at 95° for 50 sec, annealing at 60° for 50 sec, extension at 72° for 85 sec; and then terminated at 72° for 10 min. A water bath thermocycler manufactured at Pioneer Hi-Bred International was used for PCR reactions. PCR products were prepared for electrophoresis by diluting 3 μl of each product to a total of 27 μl using a combination of PCR products generated from other loci for that same maize genotype (multiplexing) and/or dH20. Dilution of 1.5 μl of this mixture to 5 μl with gel loading dye was performed; it was then electrophoresed at 1700 V for 1.5 hr on an ABI model 377 automated DNA sequencer equipped with GENESCAN software v. 3.0 (PE-Applied Biosystems, Foster City, CA).

PCR products were sized automatically using the "local Southern" sizing algorithm (Elder and Southern 1987). After sizing of PCR products using GeneScan, alleles were assigned using Genotyper software (PE-Applied Biosystems). Generally, allele assignations for each locus were made on the basis of histogram plots consisting of 0.5-bp bins. Breaks between the histogram plots of >1 bp were generally considered to constitute separation between allele bins; however, other criteria, such as the presence of the nontemplate-directed addition of adenine (+A addition) and naturally occurring 1-bp alleles, were used on a marker-by-marker basis to define the allele dictionary. All allele scores were made without knowing the identities of the maize genotypes.

RESULTS

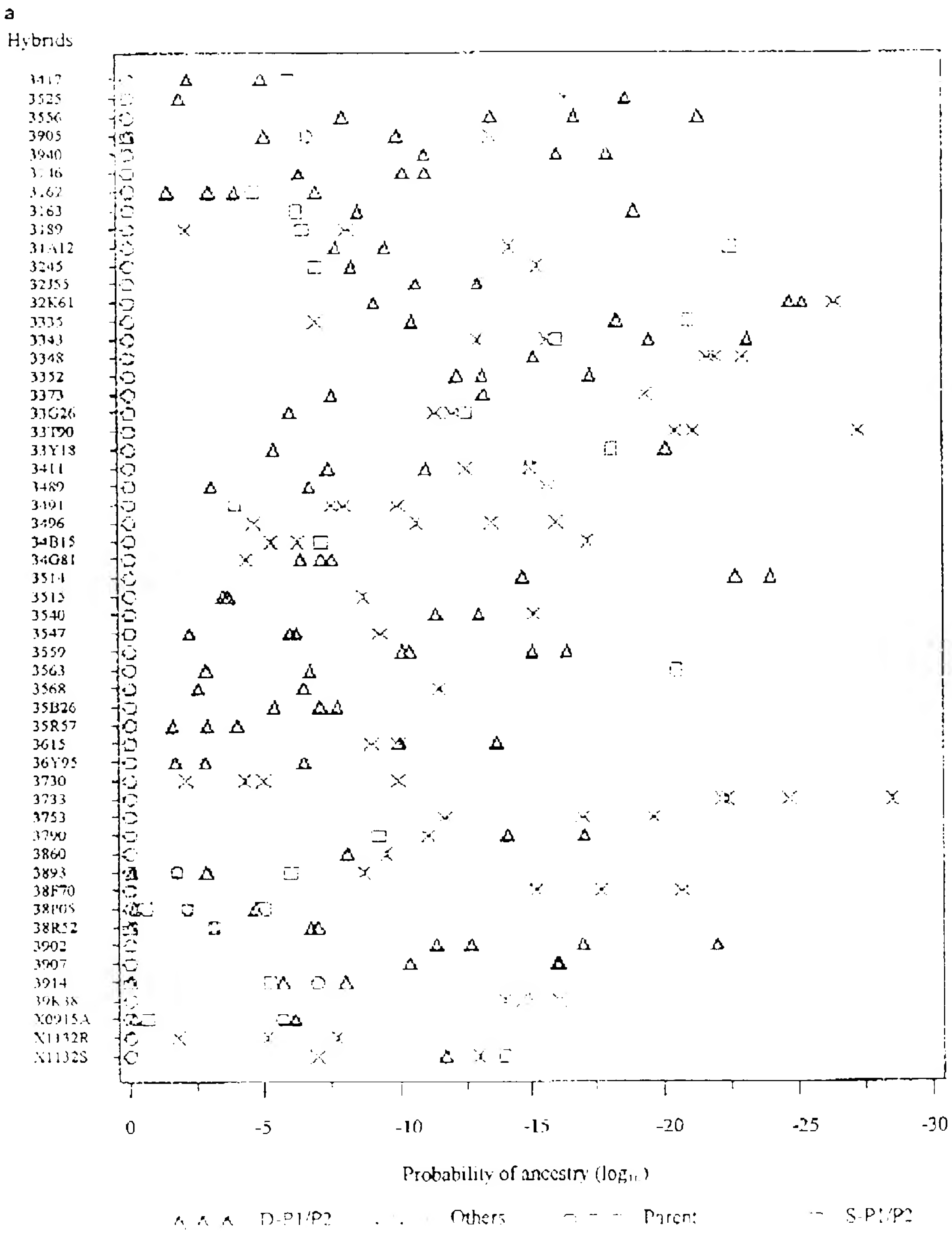
[illegible]

[illegible] and $P = 0.94$ (Figure 1b). Results for the hybrids presented in Table 2 are featured at the top of Figure 1

TABLE 2

Probability of ancestry of five hybrids using data obtained from 50, 100, and 195 SSR loci

	50 loci			100 loci			195 loci		
Hybd.	Inbd.	Prob.	SE	Inbd	Prob.	SE	Inbd.	Prob.	SE
				A. Assuming $P = 0.50$					
3417	SP1	0.9607	0.0125	P1	0.8749	0.0252	P1	1.0000	E-07
	P2	0.8077	0.1965	P2	0.8141	0.2235	P2	0.9957	0.0033
	D2P2	0.1016	0.1035	D1P2	0.1859	0.2235	D1P2	0.0043	0.0033
	D1P2	0.0907	0.0927	SP1	0.1243	0.025	D2P2	E-06	E-06
	P1	0.032	0.0125	D1P1	0.0009	0.0002	SP1	E-06	E-07
3525	P1	0.8545	E-07	P1	0.9999	<E-20	P1	1.0000	<E-20
	P2	0.8188	E-07	P2	0.5437	<E-20	P2	0.9635	0.0528
	D1P2	0.1699	E-07	D1P2	0.4563	<E-20	D1P2	0.0365	0.0528
	GP1	0.1441	E-07	CP1	E-07	E-18	SP1	E-15	<E-20
	GP2	0.0110	E-08	SP1	E-07	<E-20	GP2	E-16	<E-20
3556	P1	1.0000	E-06	P1	0.9999	E-10	P1	1.0000	<E-20
	P2	0.9616	E-08	P2	0.9997	E-10	P2	1.0000	<E-20
	D1P2	0.0340	E-10	D1P2	0.0003	E-14	D1P2	E-09	<E-20
	GP2	0.0043	E-09	D2P2	E-03	E-15	D2P2	E-11	<E-20
	D2P2	0.0069	E-10	D3P2	E-06	E-17	GGP2	E-17	E-17
3905	D1P1	0.9822	E-08	D1P1	0.9803	0.0058	P1	1.0000	E-08
	SP2	0.4927	E-07	SP2	0.6280	0.0976	D1P2	1.0000	E-06
	D2P2	0.2836	E-07	D1P2	0.2321	0.0617	D2P2	E-06	E-06
	D1P2	0.1622	E-07	D2P2	0.1317	0.0372	P2	E-07	E-13
	P2	0.0565	E-07	P1	0.0197	0.0058	D3P2	E-10	E-16
3940	P2	0.9997	0.0001	P2	0.9999	E-05	P2	1.0000	E-09
	D1P2	0.9203	0.0009	P1	0.9970	0.0011	P1	1.0000	E-09
	P1	0.0648	E-05	D1P2	0.0030	0.0011	D1P2	E-11	E-11
	D1P1	0.0127	E-05	D2P2	0.0001	E-05	DP1P2	E-17	E-17
	DP1P2	0.0014	0.0009	DP1P2	0.0001	E-07	D2P2	E-19	E-18
				B. Assuming $P = 0.99$					
3417	SP1	0.9995	0.0001	P1	0.9999	E-05	P1	0.9999	E-08
	P2	0.8836	0.1658	P2	0.9938	0.0107	P2	0.9999	E-08
	D1P2	0.0722	0.1029	D1P2	0.0061	0.0107	D1P2	E-11	E-11
	D2P2	0.0441	0.0628	D1P1	E-05	E-06	D2P2	E-14	E-14
	P1	0.0004	0.0001	SP1	E-05	0	SP1	E-20	E-21
3525	P1	0.9999	0	P1	0.9999	0	P1	1.0000	0
	P2	0.8991	0	D1P2	0.9749	0	P2	0.6135	0.4446
	D1P2	0.1008	E-11	P2	0.025	0	D1P2	0.3864	0.4446
	GP1	E-05	0	D2P2	E-20	0	GP2	E-48	0
	GP2	E-06	E-17	SP1	E-24	0	D2P2	E-49	0
3556	P1	1.0000	0	P1	1.0000	0	P1	0.9999	0
	P2	0.9996	0	P2	0.9999	0	P2	0.9999	0
	D1P2	0.0003	0	D1P2	E-09	0	D1P2	E-22	0
	D1P1	E-11	0	D3P1	E-21	0	D2P1	E-49	0
	D2P1	E-13	0	D2P1	E-21	0	D3P1	E-54	0
3905	D1P1	0.9999	0	D1P1	0.9999	E-08	P1	1.0000	E-09
	P2	0.9992	0	P2	0.9999	E-05	P2	0.9947	E-09
	SP2	0.0006	0	D1P2	E-06	E-06	D1P2	0.0052	E-11
	D1P2	E-05	0	SP2	E-07	E-13	D2P2	E-18	E-18
	D2P2	E-06	0	D2P2	E-09	E-10	D1P1	E-25	E-25
3940	P2	0.9999	E-08	P2	1.0000	E-08	P1	1.0000	E-09
	D1P2	0.9999	E-08	P1	0.9999	E-05	P2	1.0000	E-09
	P1	E-06	E-13	D1P2	E-05	E-05	D1P2	E-24	E-24
	D1P1	E-08	E-15	D2P2	E-12	E-11	DP1P2	E-41	E-41
	DP1P2	E-12	E-12	DP1P2	E-21	E-21	D2P2	E-50	E-49


a
Hybrids
3417
3525
3556
3905
3940
3146
3162
3163
3189
31A12
3245
32J55
32K61
3335
3343
3348
3352
3373
33G26
33T90
33Y18
3411
3489
3491
3496
34B15
34G81
3514
3515
3540
3547
3559
3563
3568
35B26
35R57
3615
36Y95
3730
3733
3753
3790
3860
3893
38F70
38P05
38R52
3902
3907
3914
39K38
X0915A
X1132R
X1132S
0
-5
-10
-15
-20
-25
-30
Probability of ancestry (log10)
D-P1/P2
Others
Parent
S-P1/P2

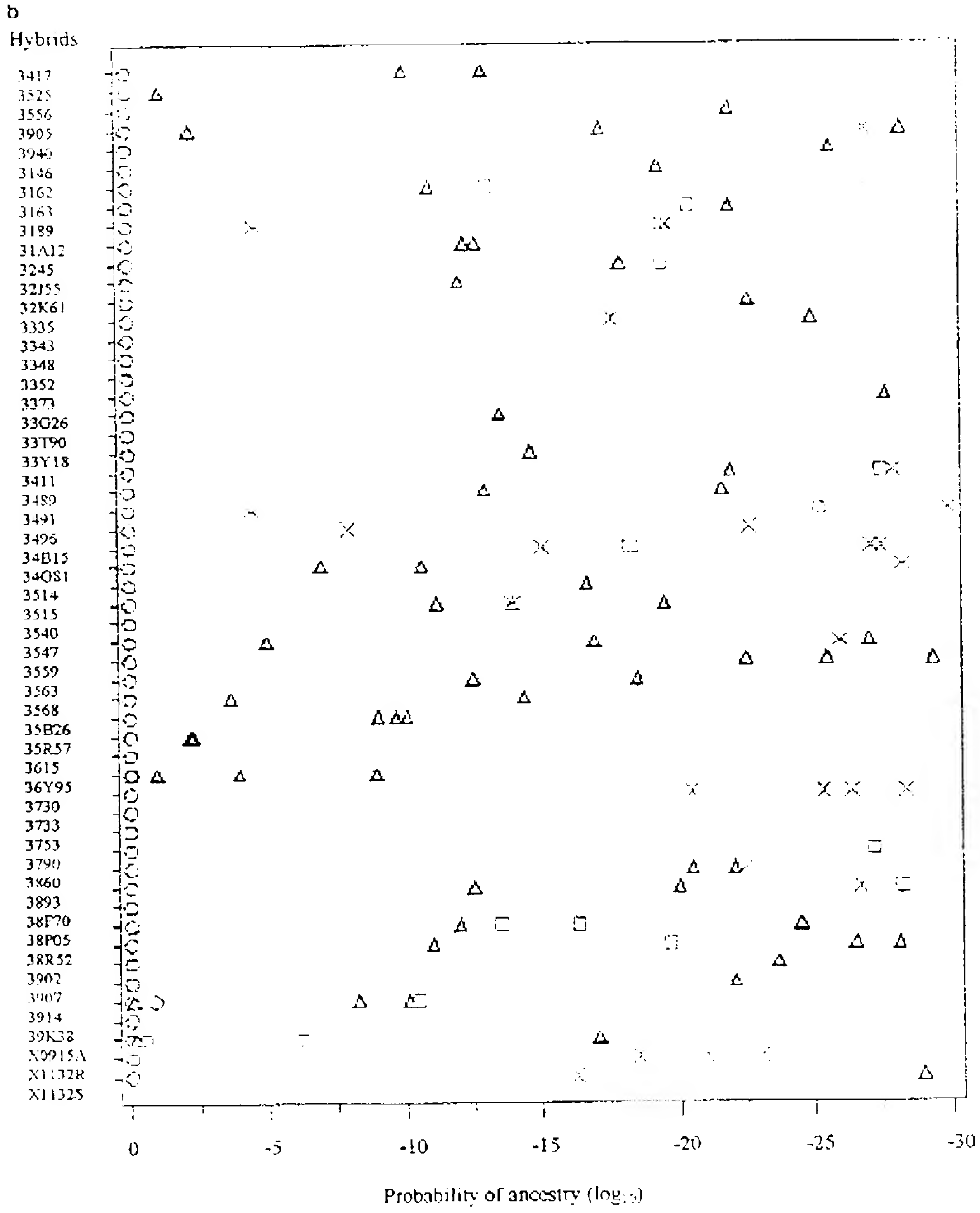


FIGURE 1.—Continued.

ranked at the top two places. The other parent was supplanted either by a sister inbred or by an inbred from [illegible] from a group of other inbreds ranged from [illegible] 0.9997, parental lines ranking second ranged from

1.0000 to 0.9653. For 35 hybrids, both parents had probabilities of ancestry in excess of 0.999. Probabilities of ancestry for nonparents that ranked in first or second places were from 0.9999 to 0.7054. For the majority of hybrids, the probability of the third and highest ranked nonparental inbred was at or below E-06. This indicates that there is usually very little uncertainty about closest ancestors.

When the algorithm used $P = 0.99$ to examine each of the 54 hybrids, both parents were correctly identified for 52 (96%) of hybrids and for 98% (102/104) of the parents across all hybrids (Figure 1). Two hybrids (3914 and X0915A), in which one parent was not ranked in the top two, were also in the subset not ranked in the top two assuming $P = 0.50$ (above). In both cases their ranks improved (both to third rank) and the actual parent was supplanted by an inbred that was a direct progeny of the corresponding parental line. For 49 hybrids, both parents had probabilities of ancestry in excess of 0.999. Among the 5 hybrids having a parent ranking second with a probability of ancestry below 0.999, the lowest of these probabilities was 0.8976 and the highest probability for a third ranking nonparent was 0.1023. For most hybrids the probability for the third and highest ranked nonparental inbred was at or below E-10.

Table 2 also addresses data analysis in circumstances where heterozygous loci occur in inbred lines or where a hybrid is scored for the presence of more than two alleles per locus. The presence of more than a single allele per locus in inbred lines is an infrequent occurrence in well-maintained inbred development and seed increase programs but is possible because ~3–5% of loci can still be segregating and unintended pollination from genotypes not designated as parents of the hybrid can occur. For hybrids, more than two alleles per locus can be scored when DNA is extracted from a bulk of individual plants and because inbred parents are not homozygous due either to residual heterozygosity or to contamination or because one or more direct parents of the hybrid are themselves hybrids. The presence of more than one allele per locus in an inbred line and more than two alleles per locus in a hybrid therefore can be accommodated by multiple runs of the algorithm, each with a random choice of two alleles per locus. Consequently, standard errors in the case of analyzing data from 195 loci tend to be very small because there were few loci where an inbred or hybrid sample (from a bulk of individual plants) was scored for more

rors. We therefore investigated the robustness of the algorithm by examining the effects of modification in the data for five hybrids (3417, 3525, 3556, 3905 and 3940). First, we reduced the number of SSRs used, from the full set of 195 to 100 and then to 50 (Table 2). Use of 50 loci generated incorrect rankings of one parent for each of two hybrids (3417 and 3940) and for both parents of one hybrid (3905). All of these most highly ranked nonparental inbreds were closely related to the true parents for each of the respective hybrids; six different inbred lines were involved. Four were direct progeny of the true parents (one with additional backcrosses from the true parent) and two were full sisters (from a cross of highly related inbreds) of the actual parent of the hybrid. Using 100 loci resulted in correct parental rankings for all hybrids except for 3905 where neither parent ranked in first or second place. Four inbreds outranked the true parents of 3905. All four nonparents were closely related to the respective true parents; three were direct progeny of the true parent of the hybrid (one with additional backcrossing to that parent) and one was a full sister of the true parent. Use of data from all 195 loci corrected the placement for one of the parents of hybrid 3905. Two inbreds that were not parents of this hybrid remained ranked more highly than one of the true parents. Both were direct progeny of that parent, and one of these inbreds had additional backcrossing to that parent in its pedigree.

To address the consequences of laboratory and other sources of error, we artificially compromised data quality beyond the level originally provided by eliminating specific proportions of alleles that had been scored (establishing scenarios where various numbers of SSR alleles were not scored) and by misscoring other alleles (establishing scenarios where various numbers of SSR alleles were scored incorrectly). We also combined the scenarios of missing data and wrongly scored data. Table 3 contains a summary of the results of making these modifications in the data. For all modifications we used data from all SSR loci and we also randomly chose SSR loci to create subsets of 50 and 100 loci. In each case, the program was run 20 times for each hybrid/set of loci. When all 195 loci were examined, replications differed only according to the particular choice of alleles for loci where more than two alleles had been scored.

To evaluate robustness in the face of missing data or mistyped data, we simulated individual and combined categories of these data in the hybrid and all inbred lines at levels of 2, 5, 10, and 25% of the alleles for each of five hybrids and all inbreds beyond the level of error as originally scored by the laboratory. We examined the effects of these levels and types of error for three sizes of database: 50 loci, 100 loci, and all 195 scored loci

[illegible] used.

Examples of robustness in the face of additional error

TABLE 3

Number of parents ranked in first and second positions (maximum is 2)

		Hybrid															
		3417			3525			3556			3905			3940			
% level change	No. of loci	50	100	195	50	100	195	50	100	195	50	100	195	50	100	195	Mean % max
0		1	2	2	2	2	2	2	2	2	0	0	1	1	2	2	77
2		1	2	2	2	1	2	2	2	2	0	0	1	1	2	2	73
5		1	2	2	2	2	2	2	2	2	0	0	1	1	2	2	77
10		1	2	2	2	2	2	2	2	2	0	0	1	1	2	2	77
25		0	2	2	1	1	1	2	2	2	0	0	1	1	1	1	57
Mean % max.		40	100	100	90	80	90	100	100	100	0	0	50	50	90	90	
0		1	2	2	2	2	2	2	2	2	0	0	1	1	2	2	77
2		1	2	2	2	2	2	2	2	2	0	0	1	1	2	2	77
5		1	2	2	2	1	2	2	2	2	0	0	1	1	2	2	73
10		1	1	2	2	2	2	2	2	2	0	0	1	1	2	2	73
25		1	0	2	1	1	2	1	2	2	0	1	1	1	2	2	63
Mean % max.		50	70	100	90	80	100	90	100	100	0	10	50	50	100	100	
0		1	2	2	2	2	2	2	2	2	0	0	1	1	2	2	77
2		1	2	2	2	2	2	2	2	2	0	0	1	1	2	2	77
5		1	2	2	2	1	1	2	2	2	0	0	1	1	2	2	70
10		1	2	2	1	2	2	2	2	2	0	1	1	1	2	2	7
25		0	1	2	0	0	0	0	0	0	0	0	1	0	0	1	17
Mean % max.		40	90	100	70	70	70	80	80	80	0	10	50	40	80	90	
		43	87	100	83	77	87	90	93	93	0	7	50	47	90	93	

ered are the same as those in Table 2.

for five hybrids using subsets of 50 and 100 loci and all loci are shown in Table 3 where numbers of parents ranking into the top two places are presented. Degradation in the preferential ranking of parent inbreds at a level of 25% additional missing data was shown for one hybrid (3525) with usage of 50, 100, or all SSR loci. Degradation in the preferential ranking of parent inbreds at a level of 25% additional misscored data was shown for hybrid 3556. When both additional levels of missing and misscored data were simulated, degradation in the ability to preferentially rank inbred parents occurred for all hybrids and for all sets of SSR (50, 100, and 195 loci) except for hybrid 3417 when data from 195 SSR loci were used. Over all five hybrids, use of 100 loci improved robustness from the use of 50 loci; use of 195 loci further improved robustness for four hybrids (3417, 3525, 3905, and 3940). The degree of improvement was small, except for hybrid 3905.

We also ranked inbreds according to their probability of ancestry of hybrids when both parents and all inbred derivatives and full-sister inbreds of the respective inbred parents for each hybrid were excluded from the analysis. The results are too voluminous to present here but can be summarized as follows: Using $P = 0.50$, a grandparent of each respective hybrid ranked into first place for 41 (76%) hybrids; probabilities ranged from 0.4976 to 1.0 and most were above 0.9999. Other classes of inbreds that ranked in first position for probability of ancestry were inbreds derived directly by pedigree from a grandparent of the respective hybrid (DGP) for 13% of hybrids, inbreds derived directly by pedigree from a great-grandparent of the respective hybrid (DGGP) for 9% of hybrids, and one class (2% of hybrids) with an inbred ranked into first place that was directly related by pedigree to the great-great-grandparent of that hybrid. Inbreds that ranked in second position were related to the respective parents of the hybrid as follows. Thirty-one (57% of hybrids) were a grandparent of the respective hybrid, 11 (20%) were classed as DGP, 7 (13%) were DGGP, 1 (2%) was class DGGGP, and 4 (7%) were a great-grandparent (GGP) of the respective hybrid. Over all hybrids, two of the four grandparents ranked into first and second positions for 23 (43% of hybrids); three grandparents ranked into the first three positions for 5 (9% of hybrids). There were no instances where all four grandparents ranked into the first four positions. Thirty hybrids had a grandparent ranked into first position using $P = 0.99$. The number of grandparents ranked into the top five posi-

DISCUSSION

The prevalent use of paternity indices demonstrates that it is advantageous to have explicit probabilities of ancestry to distinguish among different pedigrees. Molecular marker profiles are rapidly becoming more extensive and cost effective to generate. Features that would advance the statistical analysis of molecular marker data to provide explicit probabilities of ancestry include the ability to calculate probabilities of ancestry where there is no *a priori* information as to the identity of one (usually the maternal) parent and robustness in the face of laboratory error.

Maize inbred lines and hybrids provide a very exacting set of materials for evaluating the discriminatory abilities of molecular data and statistical procedures that are employed to interpret those data. Hundreds of maize inbred lines of known pedigree together encompass a great diversity and complexity of pedigree relationships. Some inbred lines can be very highly related and genetically similar due to their derivation from common parentage including from parents that are themselves highly related. Consequently, relationship categories such as "sister" or "parent" when applied to maize inbreds usually refer to closer degrees of pedigree relationship and, thus, of germplasm and molecular marker profile similarity than those of the equivalently named classes of relationship for animal species. Most maize hybrids that are widely used in the United States today are constructed from pairs of inbred lines that are unrelated by pedigree, each inbred parent having been bred from a separate "pool" of germplasm. Various degrees of relatedness are possible between hybrids according to the pedigree relationships among their constituent inbred parents.

Using $P = 0.99$ in the algorithm is more specific for identifying parents than using $P = 0.50$. However, $P = 0.99$ is less robust for identifying other relatives, such as grandparents. When the algorithm was run at $P = 0.50$ there were 6 hybrids for which one parent did not rank among the top two most probable genotypes. For the remaining 48 hybrids the correct parents were identified even in circumstances where other candidate inbreds included not only full-sister lines bred from related parents but also inbreds even more closely related to the true parent by virtue of being backcross conversions of the inbred parent of the hybrid. For each of the 6 hybrids where a nonparent ranked above a true parent, that higher ranked inbred was always either a sister or progeny of the outranked true parent. The

first two positions was 0.928 (SD = [illegible]) when $P = 0.50$, and 0.9980 (SD = 0.010) when $P = 0.99$.

from [illegible] A [illegible] imates a relationship between inbred A and A' where

inbred A′ has been bred from a cross of inbreds A and B with between one and two additional backcrosses of the parental inbred A. A Malécot coefficient of relationship of 0.9680 closely approximates a relationship between inbreds A and A″ where four additional backcrosses of parental inbred A follow the initial cross of inbreds A and B.

Running the algorithm at $P = 0.99$ in comparison to $P = 0.50$ raises the probability of ancestry for the parents while diminishing the probabilities for the third and lower ranking candidate inbred lines. Use of the algorithm at $P = 0.99$ increased both the percentage of hybrids with both parents ranked in the first two positions (from 89 to 96%) and the percentage of parental inbreds that were ranked first and second (from 94 to 98%). Two hybrids (3914 and X0915A) did not have both parents ranked first and second when the algorithm was run at $P = 0.99$. For both of these hybrids the nonparental inbred that outranked the true parent was itself a product by pedigree from the true parent that had been created by an additional four backcrosses of that parent; the Malécot coefficient of relationship between the parent of the hybrid and the inbred that outranked that parent for these two hybrids was 0.9636.

Robustness was tested by evaluating the effects of using data from different numbers of loci and by simulating additional levels of missing and misscored data up to combined levels of 25% error beyond that which was provided by the laboratory. From our experience, error rates of 5 to 10% can occur in SSR profiling of maize due chiefly to the combined effects of residual heterozygosity among seed lots and by deficiencies in the scoring of heterozygotes in hybrids. The additional levels of simulated error, therefore, include values (up to ~35% total error) that are well outside of our experience. For five hybrids that were examined, increasing the number of loci from 50 to 100 (with no additional missing or misscored data) did reduce the number of instances where inbreds that were not parents of a hybrid outranked the true parent from four to one. Nonetheless, all of these more highly ranked inbreds, although they were not themselves the true parents of the respective hybrid, were either direct progeny or full sisters of the true parent (Table 2). Consequently, if such degrees of error can be tolerated in respect of pedigrees for inbreds that are identified as parents of hybrids, then SSR data from 50 loci of equivalent discrimination ability are sufficient. Use of data from 50 loci also evidenced robustness in the face of up to 10% additional levels of either missing or misscored data, no degradation in the [illegible] levels of error; Table 3). Use of data from 195 loci provided greater resiliency against additional levels of error. However, use of data from 195 loci was unable to provide resiliency against the negative effects of adding combined levels (at 25%) of both missing and misscored data (Table 3). At the 25% level of additional poor data integrity, inbreds that were not related to the true parent of the hybrid outranked the true parent for four of the five hybrids. Levels of missing or misscored data should, therefore, be kept below 15–20% (assuming a level of 5–10% error in the data we analyzed prior to simulating additional error).

We have previously examined the pedigrees of inbreds that are ranked into the first two positions when the true parents are removed from the list of candidate inbred lines. Usually, direct progeny or full sisters of the true parents then rank most highly (data not presented). We therefore examined the rankings of inbreds with respect to their ranking and probability of inclusion in the ancestry of each hybrid after the removal, not only of the true parents, but also of the progeny of the true parents and any full sisters of the true parents. In these circumstances the grandparents of the hybrids are ranked predominantly into top positions. Using $P = 0.50$, a grandparent ranked into first position for 76% hybrids and into second position for 57% hybrids; with $P = 0.99$ a grandparent ranked into first place in 56% of hybrids. At $P = 0.50$ two grandparents ranked into first and second positions for 43% hybrids and into the first three positions for an additional 9% hybrids. Most of the remaining inbreds that ranked into the top two positions were progeny of the grandparent. A total of 108 grandparents ranked into the top five positions when $P = 0.50$; 93 ranked into these positions when $P = 0.99$. Seventy-one grandparents ranked into the top two positions when $P = 0.50$; 55 grandparents ranked into these positions when $P = 0.99$. The mean probability of a grandparent in the top two positions was 0.9288 (SD 0.1454) when $P = 0.50$ and 0.9980 (SD 0.0104) when $P = 0.99$. Our algorithm was written to identify pairs of ancestors; alternative algorithms could be tailored to identify all grandparents once parents had been identified and removed from the list of candidate inbreds.

We have demonstrated the capability and robustness of an algorithm that can be used to show probability of parentage in circumstances where the *a priori* pedigree identity of neither parent is known. Exclusions are taken into account, thereby allowing parentage to be shown even when the two parents are not represented in the database of molecular profiles that are examined. Her-

[illegible]

in the species of interest. Using $P = 0.99$ compared to $P = 0.50$ preferentially identified more true parents and with a greater difference of probability to third placed nonparents. If there is reasonable assurance that the parents are among the candidate list of inbreds, then $P = 0.99$ should be used, if greater robustness is required, then $P = 0.50$ should be used.

Applications of our algorithm include the identification of pedigrees among individuals of plant or animal species where molecular profile datasets exist that can be interpreted in terms of segregating alleles at individual marker loci and that provide a sufficient power of discrimination. Capabilities to generate large datasets of suitable molecular profile data are already available and are increasing rapidly with the advent of single nucleotide polymorphisms. One further application of our algorithm is to assist in the protection of intellectual property that is obtained on plant varieties or upon specific dams or sires of animals through the determination of pedigrees.

LITERATURE CITED

Alderson, G. W., H. L. Gibbs and S. G. Sealy, 1999 Parentage and kinship studies in an obligate brood parasitic bird, the brown-headed cowbird (*Molothrus ater*), using microsatellite DNA markers. J. Hered. 90: 182–190.

Bein, G., B. Driller, M. Schürmann, P. M. Schneider and H. Kirchner, 1998 Pseudo-exclusion from paternity due to maternal uniparental disomy 16. Int. J. Leg. Med. 111: 328–330.

Berry, D. A., 1991 Inferences using DNA profiling in forensic identification and paternity cases (with discussion). Stat. Sci. 6: 175–205.

Berry, D. A., 1996 *Statistics: A Bayesian Perspective.* Duxbury Press, Belmont, CA.

Berry, D. A., and S. Geisser, 1986 Inferences in cases of disputed paternity, pp. 353–382 in *Statistics and the Law*, edited by M. H. DeGroot, S. E. Fienberg and J. K. Kadane. Wiley Publishing, New York.

Bockel, B., P. Nürnberg and M. Krawczak, 1992 Likelihoods of multilocus DNA fingerprints in extended families. Am. J. Hum. Genet. 51: 554–561.

Bowers, J. E. and C. P. Meredith, 1997 The parentage of a classic wine grape, Cabernet Sauvignon. Nat. Genet. 16: 84–87.

Chakraborty, R., T. R. Meagher and P. E. Smouse, 1988 Parentage analysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity. Genetics 118: 527–536.

Chakraborty, R., L. Jin and Y. Zhong, 1994 Paternity evaluation in cases lacking a mother and nondetectable alleles. Int. J. Leg. Med. 107: 127–131.

Devlin, B., K. Roeder and N. C. Ellstrand, 1988 Fractional paternity assignment: theoretical development and comparison to other methods. Theor. Appl. Genet. 76: 369–380.

Elder, J. K., and E. M. Southern, 1987 Computer-aided analysis of one dimensional restriction fragment gels, pp. 165–172 in *Nucleic Acid and Protein Sequence Analysis—A Practical Approach*, edited by M. J. Bishop and C. J. Rawlings. IRL Press, Oxford.

Ellstrand, N. C., 1984 Multiple paternity within the fruits of the wild radish, *Raphanus sativus*. Am. Nat. 123: 819–828.

Gotz, K., and G. Thaller, 1998 Assignment of individuals to populations using microsatellites. J. Anim. Breed. Genet. 115: 53–61.

Gunn, P. R., K. Trueman, P. Stapleton and D. B. Klarkowski, 1997 DNA analysis in disputed parentage: the occurrence of two apparently false exclusions of paternity, both at short tandem repeat (STR) loci, in the one child. Electrophoresis 18: 1650–1652.

Hamrick, J. L., and A. Schnabel, 1985 Understanding the genetic structure of plant populations: some old problems and a new approach, pp. 50–70 in *Population Genetics in Forestry*, edited by H. R. Gregorius. Springer-Verlag, Heidelberg, Germany.

Helminen, P., V. Johnsson, C. Ehnholm and L. Peltonen, 1991 Proving paternity of children with deceased fathers. Hum. Genet. 87: 657–660.

Lang, J. W., R. K. Aggarwal, K. C. Majumdar and L. Singh, 1993 Individualization and estimation of relatedness in crocodilians by DNA fingerprinting with a Bkm-derived probe. Mol. Gen. Genet. 238(1–2): 49–58.

Malécot, G., 1948 *Les Mathématiques de l'Hérédité.* Masson et Cie, Paris.

Marshall, T. C., J. Slate, L. E. B. Kruuk and J. M. Pemberton, 1998 Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. 7: 639–655.

Meagher, T. R., 1986 Analysis of paternity within a natural population of *Chamaelirium luteum* (L). Identification of most-likely male parents. Am. Nat. 128: 199–215.

Meagher, T. R., and E. Thompson, 1986 The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. Theor. Popul. Biol. 29: 87–106.

Miller, P. S., 1995 Selective breeding programs for rare alleles: examples from the Przewalski's horse and California Condor pedigrees. Conserv. Biol. 9: 1262–1273.

Primmer, C. R., M. T. Koskinen and J. Piironen, 2000 The one that did not get away: individual assignment using microsatellite data detects a case of fishing competition fraud. Proc. R. Soc. Lond. B Biol. Sci. 267: 1699–1704.

Rannala, B., and J. L. Mountain, 1997 Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. USA 94: 9197–9201.

Sefc, K. M., H. Steinkellner, J. Glössl, S. Kampfer and F. Regner, 1998 Reconstruction of a grapevine pedigree by microsatellite analysis. Theor. Appl. Genet. 97: 227–231.

Smith, J. S. C., E. C. L. Chin, H. Shu, O. S. Smith, S. J. Wall *et al.*, 1997 An evaluation of the utility of SSR loci as molecular markers in maize (*Zea mays* L.): comparisons with data from RFLPs and pedigree. Theor. Appl. Genet. 95: 163–173.

Smouse, P. E., and T. R. Meagher, 1994 Genetic analysis of male reproductive contributions in *Chamaelirium luteum* (L.) gray (Liliaceae). Genetics 136: 313–322.

Taylor, A. C., A. Horsup, C. N. Johnson, P. Sunnucks and B. Sherwin, 1997 Relatedness structure detected by microsatellite analysis and attempted pedigree reconstruction in an endangered marsupial, the northern hairy-nosed wombat *Lasiorhinus krefftii*. Mol. Ecol. 6: 9–19.

Thompson, E., and T. R. Meagher, 1987 Parental and sib likelihoods in genealogy reconstruction. Biometrics 43: 585–600.

Vankan, D. M., and M. J. Faddy, 1999 Estimations of the efficacy and reliability of paternity assignments from DNA microsatellite analysis of multiple-sire matings. Anim. Genet. 30: 355–361.

Wetton, J. H., D. T. Parkin and R. E. Carter, 1992 The use of genetic markers for parentage analysis in *Passer domesticus* (house sparrows). Heredity 69: 243–254.

White, E., J. Hunter, C. Dubetz, R. Brost, A. Bratton *et al.*, 2000 Microsatellite markers for individual tree genotyping: application in forest crime prosecutions. J. Chem. Technol. Biotechnol. 75: 923–926.

Communicating editor: Z-B. Zeng